Assessment of the Accuracy of GeoNames Gazetteer Data

Dirk Ahlers
NTNU – Norwegian University of Science and Technology
Trondheim, Norway
dirk.ahlers@idi.ntnu.no

ABSTRACT

Gazetteers are the basis of many geospatial applications and serve an important role to collect and make available knowledge about the physical world such as place names and their coordinates. GeoNames is one of the largest and most often used gazetteer and it is generally assumed to be of sufficient quality. In this paper, we examine the quality and accuracy of the data in more detail, triggered by some anomalies encountered during its use. We present a classification of inaccuracies ranging from grid patterns, imprecise coordinates, overlaps and repetitions as well as misclassifications and visualize these for a range of countries. We finally give an outlook into potential corrections.

Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Information Storage and Retrieval—*Information Search and Retrieval*

General Terms

Experimentation, Measurement

Keywords

Geocoding, Gazetteer, source integration, positional accuracy, entity reconciliation, data quality

1. INTRODUCTION

The basis of most approaches in geographic information retrieval and many location-based services is a database of known placenames such as cities or villages along with their geographic coordinates. Such a gazetteer [13] can additionally contain a geographical hierarchy, population numbers, translations, and other geographic knowledge about a wide range of geographic and geological entities. Gazetteers are used for grounding of placenames, ambiguity resolution, toponym disambiguation, entity resolution, geoparsing, classification, and many other tasks. The most popular freely available gazetteer is GeoNames[1], which is used in a huge number of research projects and commercial systems.

We had initially detected some anomalies in the dataset when we investigated its use as a ground truth for a search engine project in Honduras [2]. When trying to match Wikipedia articles to the gazetteer data there occurred similar or even identically names places close to the given coordinate (cf. Fig. 13). We had previously explored a similar merging for Wikipedia articles in different languages [4]. In trying to diversify, we assessed other gazetteers, but the main alternative, the Getty Thesaurus of Geographic Names[2], had only a minimal number of entries for Honduras. Its available 218 entries and 119 inhabited places are too few compared to more than 18600 entries and 12000 populated places in GeoNames.

Before we continue the assessment, we will briefly clarify some requirements for the gazetteer data.

Accurate positions should foremost reflect the location of entities in the real world. This is slightly complicated by the problem of the choice of a point representation for spatially extended areas. However, this is what is available, even if we will later discuss potential alternatives. In this case, we reformulate the problem to have the point match closely the center of an area within a certain margin of error. For points of interest, these should be at the building level, equaling an order of magnitude of 10m-50m [5] For example, neighborhoods at the sub-minute scale often have a diameter of less than 1km and should be exact to 100m-1km. For larger areas such as cities or counties, the accuracy can be weakened a bit, but for smaller villages, 1km would be expected. A weaker requirement would be that a point does not lie outside the area it describes.

Distinguishability together with correct spatial relations should make sure that disctinct entities can be identified as taking up different spaces and that relations between entities and distances are useful and meaningful. This goes along with an (adequate) uniqueness that should make sure that different places can be uniquely identified by having different features, in other words, no duplicates should exist.

A *correct feature type* is necessary to distinguish different types of entities and, e.g., infer additional information.

Coverage in all places where there is something to cover should make sure that there is a uniform expectation of data, of course adapted to building or population density.

GIR'13 November 05 - 08 2013, Orlando, FL, USA
Copyright 2013 ACM 978-1-4503-2241-6/13/11
http://dx.doi.org/10.1145/2533888.2533938 ...$15.00.

[1] `http://www.geonames.org/`

[2] `http://www.getty.edu/research/tools/vocabularies/tgn/`

2. THE GEONAMES DATASET

GeoNames integrates gazetteer data from its constituent multiple sources[3] and users can edit data in a wiki-like interface. The data comes from other public and open gazetteers, which can vary in quality, scope, resolution, or age. Obviously, some merging of the sources takes place. Unfortunately, the merging process is not published, which makes a judgement of the resulting open data difficult due to missing history, provenance, and processing information.

The minimum feature set for an entity in GeoNames is the name, the coordinate as latitude and longitude, the parent administrative division, and the country. Additional information can be population data, height, alternative or translated names, or links to Wikipedia. The entities can be natural, artificial, or political; for example administrative divisions, populated places such as cities or villages, waterbodies, parks, special designation areas, points of interest or buildings, geographical entities such as mountains, islands, undersea features, or forests[4]. The data is organized in a hierarchy down from a country level. This allows the download of all data pertaining to a country[5].

3. RELATED WORK

A lot of the work on inaccuracy and uncertainty in geospatial data [9] is concerned with positional error during geocoding [7, 10] or with high-granularity positional accuracy and precision and potential conflation techniques [5, 6]. Thematic accuracy concerning the non-geospatial features of the entities is another important aspect. Furthermore, accuracy assessments have been done, sometimes with surprising results, for Flickr[11] and gazetteers have been examined for distributions and distances of ambiguous toponyms [8]. Within the GeoNames data, different types of inaccuracies are observed. We will systematically discuss and analyze them in the following to develop an understanding of their characteristics. To the best of our knowledge this is the first study of data quality in GeoNames.

4. ANOMALY ASSESSMENT

A first visual inspection takes place at two levels, first the whole country and second a detail view of an urban region. At the country level, we first plot all GeoNames places for Honduras in Fig. 1. Country borders as high-quality polygons were taken from GADM[6], the points are plotted at 60% opacity. In the following we focus on only the populated places (PPL) as these are mostly used in geocoding and placing tasks. These are shown in Fig. 2 and represent about 65% of the full dataset.

The country appears to be well covered in most regions. The lower density in the eastern region is consistent with population numbers as this is mostly rainforest and has the lowest population density in the country[7]. However, there are some satellite points around the country. Those to the North and East are mostly islands, but there are also some clear outliers. Additionally, we observe some rectangular patterns of higher density as well as additional anomalies.

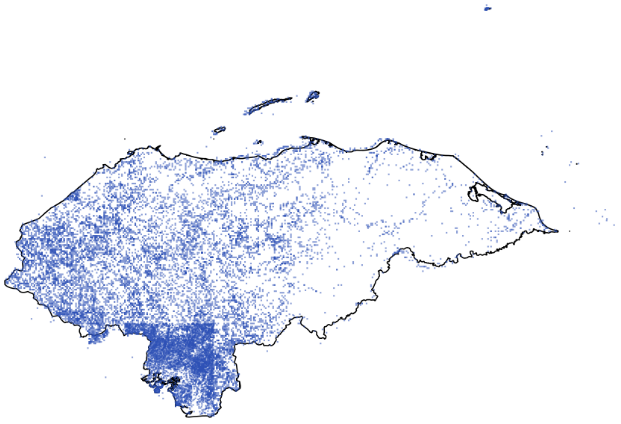


Figure 1: Plot of all geonames entries for Honduras

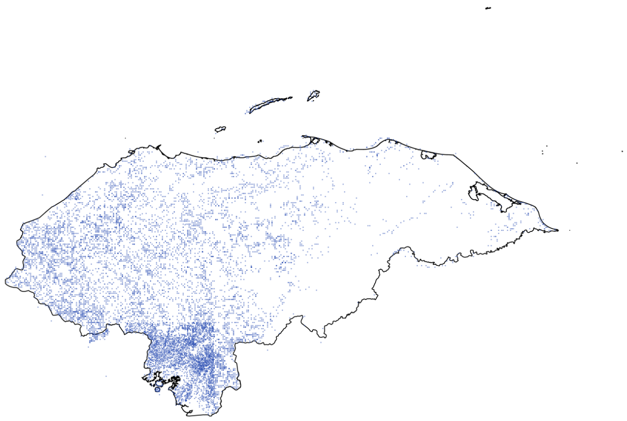


Figure 2: Plot of all populated places for Honduras

Upon closer inspection, the spatial distribution of places raised a suspicion that the places were not "naturally" geocoded. Fig. 3 shows a part of Tegucigalpa, the capital of Honduras. The grey markers are populated places, the coloured markers show entities other than populated places, for example 2 is a hotel, 13 is a misplaced hotel, 25 is an airport, 37 is a mountain ridge. 11 is the marker for the capital itself: Tegucigalpa, 14.0818, -87.20681, N 14° 04' 54" W 87° 12' 25". 1 is Florencia, a neighborhood: 14.08333, -87.18333, N 14° 05' 00" W 87° 11' 00". When hovering over some items in the map Web interface, GeoNames shows a bounding box that roughly fits the error described here. However, the box is always a square as an approximation based on a combination of features[8], it does not show the true area or the margin of error. On this view of Tegucigalpa, we can observe several inaccuracies and anomalies:

- Grid-aligned rastered positions: most places are snapped to the corner points of a rectangular grid.
- "Holes" in the coverage: despite a regular coverage along the grid, some areas contain no information. However, there are neighborhoods at the sub-minute scale and at the holes, so they should be covered.

[3] `http://www.geonames.org/data-sources.html`

[4] `http://www.geonames.org/export/codes.html`

[5] `http://download.geonames.org/export/dump/`

[6] `http://www.gadm.org/country`

[7] `http://data.worldbank.org/country/honduras`

[8] Sometimes actual bounding boxes or even polygons are available in the subscription version. `http://geonames.wordpress.com/2013/06/26/new-map-layout/`

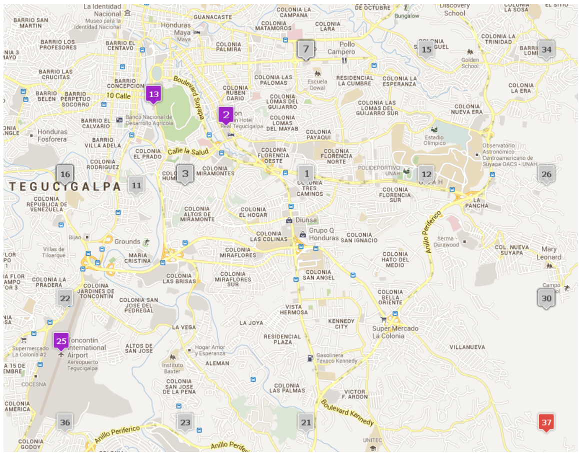


Figure 3: Detail of the map for Tegucigalpa, shown on the geonames.org map view

- Repeated positions: for example, marker 30 hides 3 more items of different places.
- Inaccurate feature classes: Florencia is a neighborhood, but is classified as a regular populated place.
- Additionally there are places outside the country as seen in Fig. 2 and also near-identical places which will be discussed later.

4.1 Grid-aligned positions

Geocoding should be based on the actual position of a place. A prominent rectangular grid pattern in the data is indeed surprising. As GeoNames exports coordinates in the decimal degree notation, the items on the grid have coordinates such as 14.08333, -87.18333. This does not directly reveal the cause of the grid. However, when converting them to the degree-minute-second notation, the pattern becomes more obvious: N $14°05'00''$ W $87°11'00''$. The grid is not arbitrary, but rather the result of the seconds omitted from the coordinates. Thus a part of the places is left with a granularity of only degrees and minutes and snapped to a graticule with spacing of $\Delta\phi = \Delta\lambda = 1'$. Due to the location of Honduras, the difference between minutes of longitude is about 1.8 km. We discuss this spatial resolution in Section 4.3.

If we assume the actual coordinates of places are determined without reference to the grid, then we should see a more or less uniform random distribution of the minutes and seconds. Certain other distributions such as Zipf's Law [14] are not expected to hold at this level. In the degree-minute-second (DMS) notation, the chance for the seconds to be all zero is only $\frac{1}{100}$. The chance of both dimensions of a coordinate being exact to the minute becomes only $\frac{1}{10000}$. This is not observable in the data, where the fraction of on-grid coordinates is much higher. Within the populated places, there are 9391 truncated on-grid and 2846 more naturally distributed non-zero-seconds full coordinates, a ratio of 77% : 23% (cf. Table 1). This means that the truncated seconds are an extremely widespread phenomenon and introduce a strong bias into the data.

To drill down, we partition the data into the truncated and full coordinates. As mentioned above, about 0.01% of the truncated data would be expected to naturally occur without seconds, but we have of course no way to identify those. First, we look only at the seconds from the coordinates. If plot them as seen in the left of Fig. 4, the general data for Honduras shows almost all coordinates' seconds hitting zero. On the right side, we only plot the remaining off-grid points, and receive a more even distribution as would be expected.

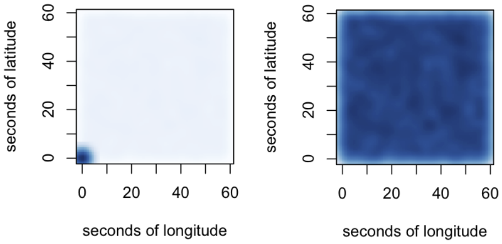


Figure 4: Density plot of the distribution of the seconds part of coordinates for (a) all points and (b) points with intact non-zero seconds

We plot these partitions of the coordinates at the country level to get a better idea how they are distributed. Fig. 5 shows imprecise coordinates in red and those with more significant digits in blue. The suspicious rectangular areas spotted earlier in the data now resolve in a surprising way. The coverage of the country is mainly effected by the off-grid data. The more precise data is heavily concentrated in a big southern and a much smaller northwestern area as well as some very sparsely distributed additional points. These seem to be true additional points, with no observable decline of the truncated data density in these regions. There is nothing observable special about these regions, including that they do not contain larger cities or the capital and are not departments. This means that a large part of the country is only covered by low-precision data, while the high-precision data comes from an unknown source of limited spatial coverage. Unfortunately, geonames does not make any provenance metadata available to identify the actual data source[9] of a given entry. So we can only speculate that the high-precision data comes from a different source that was merged into the general geonames dataset, but are not able to separate the data by source.[10] Apart from the rectangular areas, the low number of other scattered points could be the result of manual edits and corrections in the dataset, which are possible for registered users.

Since the plot does not show the details of very dense regions, we have opted for a density plot that can show the distribution within the point patterns. In Fig. 6 we can see that the density of the full coordinates is higher in the southern area than in the small northwestern one.

Such a major discrepancy between the on- and off-grid coordinates hints at some problem with the dataset. The patterns in the data also implicates that there is not a uniform expectation of coverage. Either inside the higher density regions there are superfluous or very small places or outside them, certain places or types of places are missing. Honduras being a developing country, we wanted to test whether

[9] `http://www.geonames.org/data-sources.html`

[10] It might be possible to query the individual sources, but this would defeat the purpose of geonames as a general one-stop gazetteer.

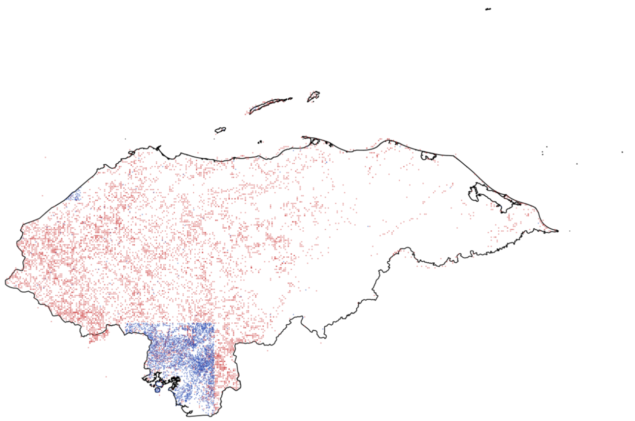


Figure 5: Plot of on-grid (red) and off-grid (blue) populated places for Honduras

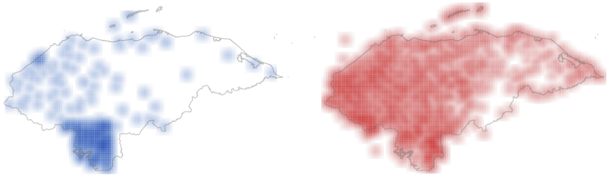


Figure 6: Smoothed density plot of the partitioned coordinates from Fig. 5

this might be a regionally limited phenomenon.

The same partition was applied to all neighboring countries of Honduras in Fig. 7. Most obvious is that the phenomenon of truncated data is more widespread and covers also El Salvador in the south and Guatemala in the west. However, for Nicaragua in the southeast, the situation is inverted and shows a much higher value of precise coordinates. The rectangles of precise data are crossing borders (with lower density), but there are also other more complex shaped patterns. Nicaragua also shows a continuation of the band of rainforest with fewer populated places. The wedge-shaped hole in its southwest is a lake with an island.

For a full overview, we examine all Central American countries in Fig. 8. We see that Nicaragua and Costa Rica both have a high ratio of precise coordinates, but that Panama shares the characteristics of the other countries with a useful coverage only given by the low-precision data.

Finally, for a comparison with industrialized countries, we have a look at Germany and Norway. The amount of truncated coordinates is much lower at around 21% for Germany, with Norway being in the middle with 56%. Norway (cf. Fig. 9) with its low population density shows large empty areas; the population centers along rivers and fjords and the capital in the south can be spotted easily. The truncated coordinates are distributed similar to the full ones. Germany (cf. Fig. 10) is very well covered, empty areas mostly correspond to mountain ranges. There are four regions of higher density for the truncated coordinates, these are Schleswig-Holstein in the north, Rhineland-Palatinate in the west, Thuringia, less obvious, in the center, and a small region in the south that does not correspond to a state. As geonames integrates datasources on all levels, this is supposed to mostly stem from respective state-level sources.

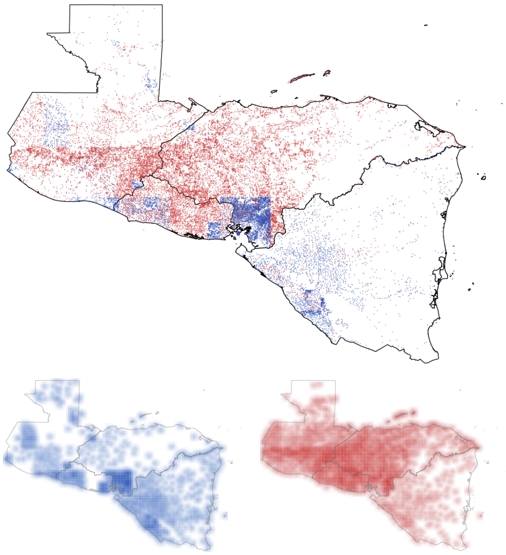


Figure 7: Neighboring countries: Guatemala, El Salvador, Honduras, Nicaragua (left to right)

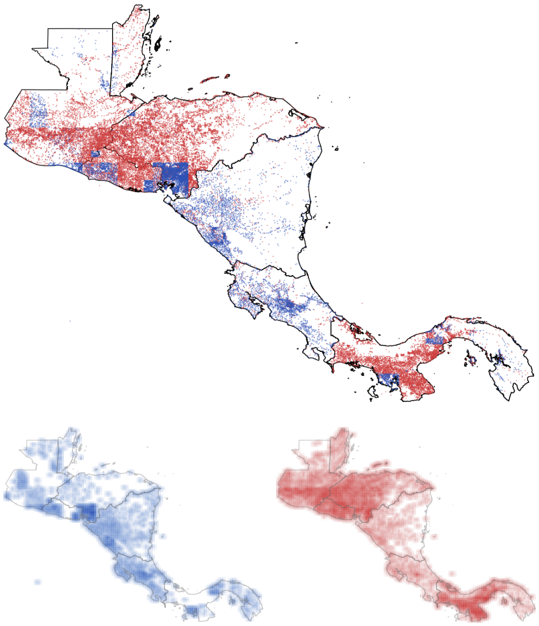


Figure 8: Central America: Guatemala, Belize, El Salvador, Honduras, Nicaragua, Costa Rica, Panama

4.2 Characteristics of grid-aligned points

The rather uniform distribution of the full grid-pattern is a curious artifact of a process that we have not yet been able to ascertain. For a possible correction, it is important to distinguish whether the on-grid coordinates were rounded or truncated to the observed low resolution.

As GeoNames would not reduce the coordinates, the low-precision data comes from the data sources. One hypothesis is that the places were generated by inverse geocoding on a

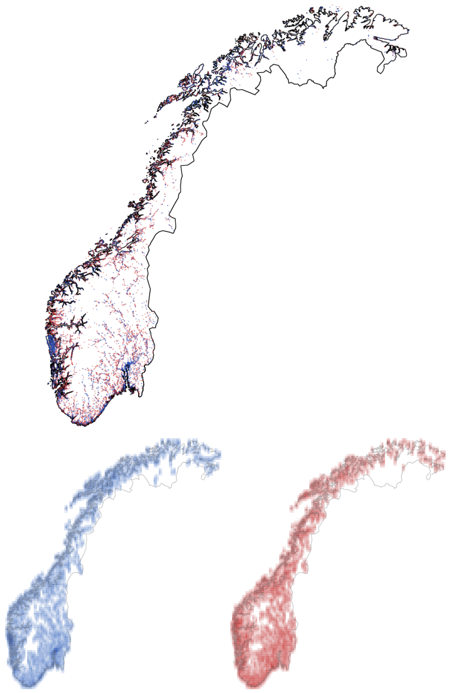


Figure 9: Scatterplot and density plot for Norway

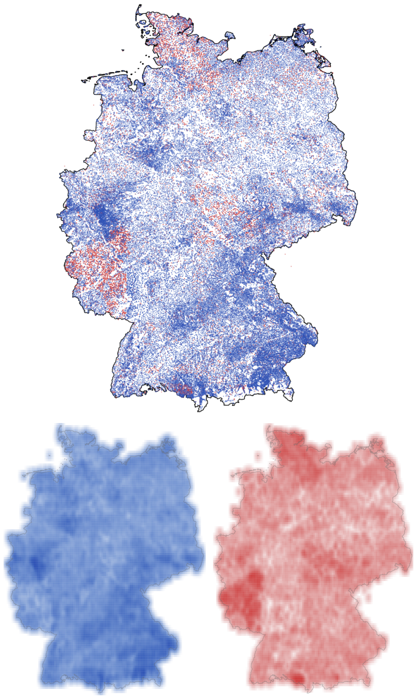


Figure 10: Scatterplot and density plot for Germany

rastering of the grid points. However, it is more probable that the data sources simply contain this reduced precision due to low-resolution geocoding. In both cases, the error distribution would rather be a rounding than a truncation. One finding that would support either the rastering hypothesis or indicate a rather bad data source is that we find some extremely small places within the raster, but also find larger places missing from the data.

In case of truncation, coordinates would be systematically shifted from the quadrant of coordinates down to the grid point. In case of rounding, places would be distributed in a rectangle with the extent of half a minute in each direction around the grid point. To confirm this, we took a sample of some places with truncated coordinates. We made a visual comparison to Google Maps satellite images (cf. Section 4.6) because we could not be sure about the name placement or the accuracy of other gazetteer data. However, even with added inaccuracies, a tendency in the shift would be detectable, unless the mapping error would be exactly by half a minute in both directions. A potential shift by a general mismatch in reference systems is unlikely, as discussed in Section 4.6.

For the examination, we left out near-identical places such as in Fig. 13, and also places that are much bigger than the expected margin of error. Some exemplary results are shown in Fig. 11. The difference between GeoNames places and actual place location shows random directions, distances within the rectangle around the grid point, but also some higher distances as also discussed in Section 4.6. No systematic shift or bias in the shortened coordinates was detectable. Therefore, no correction method presents itself based on only the GeoNames data. A solution would be instead to add error annotation to the data or to include other data sources (cf. Section 4.3).

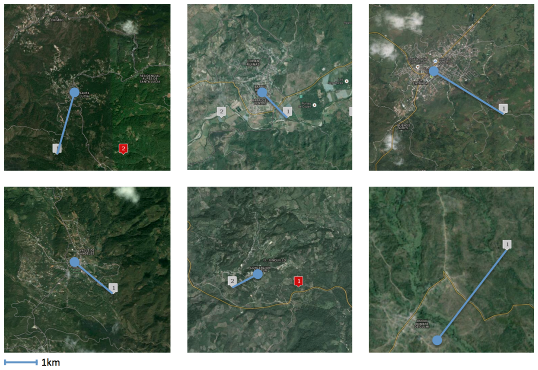


Figure 11: Direction and distance of error for places with truncated coordinates

4.3 Discussion of granularity

Granularity and precision in coordinates can be deceiving. Because the coordinates are exported with all digits, it is impossible to identify whether a coordinate was truncated or naturally occurred at a zero position. However, as discussed previously, in a uniform distribution this would only occur in 0.01% of cases. Furthermore, in the decimal notation, the zero values get concealed by the conversion. This is related to the often discussed issue of NULL in databases. It remains unclear whether the zero-value seconds are inapplicable, unknown, or actual data. A derivation of precision as the number of significant digits is therefore impossible. As a long-term solution, a scientific notation of powers of ten could reduce the number of significant digits. Alternatively, a plus-minus notation could indicate the tolerance for the measurement.

Furthermore, the truncation of seconds is not a reliable

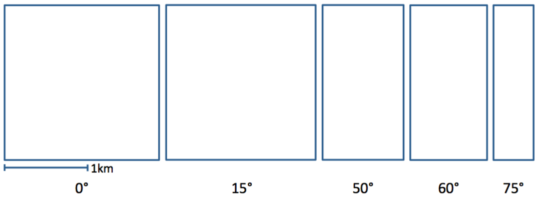


Figure 12: Margin of error dependent on latitude

method to signify reduced precision. An interesting effect is that the use of truncated coordinates does not give a uniform expectation of uncertainty on the earth. The error is not fixed, but depends on the latitude because the grid size formed by lines of latitude and longitude varies across the globe and gets narrower toward the poles.

The latitudinal error varies little around 1.85km over the globe, but the longitudinal extent varies from 1.85km at the equator to 0.47km at 75° latitude and reaches zero at 90°

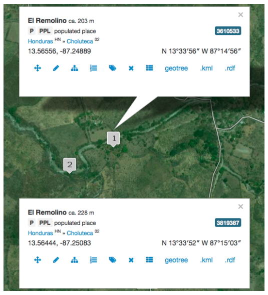
Fig. 12 shows this distribution of error sizes for selected regions. This would also be the minimal annotation of uncertainty to add to the respective places. But as we showed in before, the distances can even be higher. We expect to use such an annotation [19] of the spatial extent or footprint [18] in the future.

4.4 Inaccurate feature classes

In the map of Tegucigalpa in Fig. 3, we see some entries describing neighborhoods, yet are given as populated places. For example, the neighborhood "Florencia" has the feature code PPL (normal populated place) with the hierarchy "Francisco Morazán", "Honduras". GeoNames provides the code PPLX for a section of a populated place, e.g., a city district, but its use is not consistent. For Honduras, we find only 3 such entities, while, e.g., Norway has a more credible number of 225 or about 2.5% (cf. Table 1). The problem is that this can look like many small places instead of a large city, that georeferencing may not work properly, and that inverse geocoding may miss the actual city. One way to partially detect this is to use the hierarchy. Unfortunately, the neighborhoods are not inserted under the respective city, but have the administrative region as their parent in the geotree. However, we have found that in some cases, the element does not appear directly under the hierarchy of its parent administrative region in the geotree. Combining with the availability of population data can in some cases help to identify a neighborhood as part of a large city, but is not yet conclusive. This is in part due to the low percentage of available population numbers for places, which is between 3 and 7% for the Latin American countries, 11% for Norway, and 26% for Germany.

4.5 Near-identical places

There are some situations where multiple places have the exactly same name, the same feature code and hierarchy and are placed within a close distance of each other. Consider the case of El Remolino in Fig. 13. The two places are very close to each other, with a distance of only 0.24km. None of the places fits within the truncated grid, which would have been a helpful hint for correction. In this case, the satellite image shows a very small settlement closer to marker 1 than to marker 2. An average of the two positions would still be acceptable in this case. At this stage, we consider only identical names, the same feature code, and a maximum distance of 2km. Name variations for nearby places are part of our future work on entity merging. Using this method, we find 232 repeated places for Honduras. These are found in clusters of varying size. The most are found in 86 clusters of size 2, which translates to 172 affected places, but the cluster size goes up to 6 (cf. Fig. 14).


El Remolino ca. 203 m
P PPL populated place 3610533
Honduras HN » Choluteca 02
13.56556, -87.24889 N 13°33′56″ W 87°14′56″

El Remolino ca. 228 m
P PPL populated place 3819387
Honduras HN » Choluteca 02
13.56444, -87.25083 N 13°33′52″ W 87°15′03″

Figure 13: Example of near duplicates differing only by coordinate

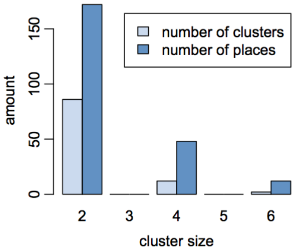


Figure 14: Histogram of near duplicates, number of spatial clusters and affected places

In certain cases, a populated place (PPL) shares the exact coordinate with a place of the same name, but the PPLA or ADM feature class. This can be explicitly disambiguated because this means that the department share the name with its capital and both were plotted at the exact same coordinates. Identical places hardly occur, as these should be unified by the internal GeoNames merging process. Only Norway exhibited this with 6 affected places. Other cases of repeated places concern feature types such as rivers or other geological entities, which were found to sometimes have their path plotted in the data.

Due to the grid alignment, there is another widespread phenomenon where places share their coordinate with one or more other places, with the same feature code, but differing in name. Such repeated coordinates for places can break topology and distinguishability of places. There are 4823 affected places in Honduras. The data for all countries is listed in Table 1.

4.6 Places outside their assigned country

As described earlier, we retrieve the entities by the grouped dumps that are available for each country and comprise all entities that have this country assigned. When plotting them over the maps from GADM, we see some points outside the boundary of the country's landmass. These fall in roughly two categories, points close by the border or coastline, and faraway points.

Looking back at Fig. 1, we see some points at the Caribbean coast in the North, south of the islands, that are moved into the ocean, and seem to follow a shifted coastline. There are also other satellite points all around at sometimes rather high distances and a continuous area in the southwest outside the border. The effect is clearly reduced, but not removed, in Fig. 2, which shows only populated places, other features such as reefs, waterbodies etc. were filtered.

Following up on the latter case, Fig. 15 shows a zoomed detail of the border between Honduras and El Salvador. In this map, all places are plotted, the color signifies the country. There is a focused overspill of places assigned to Hon-

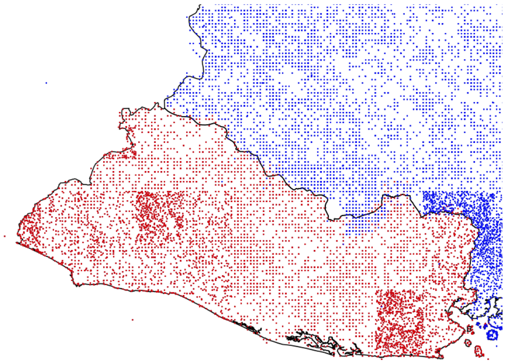


Figure 15: Zoomed plot of all places for Honduras (blue) and El Salvador (red) showing overlap and empty regions

duras (blue) into the region of El Salvador, which has no places in that area. There are also areas of mixed occurrence of places from both countries and a completely empty region. These effects could be traced to be the leftover of an old conflict from a war in 1969 that left some disputed areas and unsettled borders, so-called bolsones, that were only settled in 1998[11] and have obviously not yet been reflected in the data. On the other hand, we have not yet found an explanation for the overspill in the northeast. The figure also shows some faraway outliers. In general, empty regions are less of a problem because they correspond in most cases to areas of lower population density such as rainforest, glaciers, or mountain ranges. Far distanced satellite points that fall on other territory were not found to exhibit clear characteristics.

While borders and territorial waters may suffer from political affairs, coastal lines do not and may be better suited to estimate errors. For example, in Fig. 2 the Caribbean islands to the north appear to be partially shifted. Their points are off to the west, while the coast directly south of them has some points, judging from the coastline repeated in the point patterns, moved to the northwest. Based on some known cities, we can estimate the shift as roughly west-northwest by about 5.5km. These places themselves have a size of about 1km. Yet, some of the places in these areas are correctly placed (within the limits of the grid) to the actual position of the town. However, the islands in the South of the country appear to be targeted very well. At a latitude between N 12.9 and N 17.4, we cannot explain this by a consistent projection or referencing error.

One possible explanation is that some of the data was plotted using a different reference system that systematically skewed the coordinates. However, this does not seem to be universal, even when judging each data partition individually. We have currently no method to separate the two classes. If we would only reposition those in the ocean, we would potentially fold them over some points with that error that are on the landmass, and thus worsen relative positions and topography.

We are aware that these comparisons could be flawed by comparing to incorrect data. However, we aimed to reduce the potential influence by using multiple sources. We checked on our own extracted data, and also directly on the map interface provided by GeoNames[12], which uses Google Maps. These, as well as the GADM polygons for the plotted borders, were compared with Google, Yahoo! and Bing Maps satellite image data. These generally agree. Otherwise, the reported offset for these data sources based on aerial imagery is generally reported below 50m. E.g., [16] reports a difference between developed and developing countries with around 25m and 45m root mean squared error, respectively. These values are much lower than the potential error on the data, which is in the range of kilometers. Of course, we still have to be careful as those datasources are of unknown quality. While in developed countries, the data is mostly correct (but still of unknown quality), accuracy or even availability can often be reduced in developing countries.

Determining the selected areas affected by a potential skewing appears to be very difficult to identify automatically as this seems only viable as a pattern analysis on a larger set of points, not for individual places. As the underlying datasource is not identifiable from the data, this deprives us of the most promising feature for correction.

4.7 Quality indicators

The discussed inaccuracies can serve as initial quality indicators. They are aggregated in Table 1 for the different countries examined here. We expect to extend this to more countries and also aim to compute an indicator number or a measure of the margin of error as an inaccuracy annotation. The table first collects for each country the number of populated places, the amount of full and shortened coordinates. It then shows the percentage of shortened coordinates, of near-identical repeated places and of overlapping places. For both it counts the number of affected places, i.e., if two places overlap, the number will be 2. These numbers would ideally be low. PPLX gives the amount of places classified as parts of cities, e.g., neighborhoods. Its expectancy depends on the number and size of cities per country. The countries are grouped by their region, first the Central American countries, followed by the European countries for comparison.

5. CONCLUSION AND FUTURE WORK

The aim of this work was to assess the precision and accuracy in the GeoNames gazetteer data. The analysis has shown that there exist inaccuracies of various types and given partial error estimations. The data shows some random as well as systematical errors which in some cases appear indistinguishable from correct data. For example, outliers outside the country can be seen, but their continuation into the rest of the dataset could not be determined. A very interesting finding is that apart from other issues, shortened coordinates occur in very high numbers. In many cases, certain areas have an extent at the sub-minute scale that is not captured. Compare for example the neighborhoods in Fig. 3. On the other hand, we see that the capital (marker 11) is moved off the grid and has a full coordinat even though it has a larger extent than the other places. Also, it damages topology [1], making spatial relations of places to each other less useful, demonstrated by high numbers of overlapping places (cf. Section 4.5, Table 1). Inaccurate feature types

[11] http://www.ipgh.org/download-file/boletin-aereo/Boundary.pdf

[12] http://www.geonames.org/maps/

Table 1: Quality indicators for the examined countries

		coordinates			places				
Country	PPL	full	shortened	% shortened	repeated	% repeated	overlapping	% overlapping	PPLX
Honduras	12237	2846	9391	77	232	1.9	4823	39	3
El Salvador	3844	1074	2770	72	54	1.4	437	11	26
Nicaragua	2938	2394	544	19	14	0.5	8	0	57
Guatemala	6497	939	5558	86	46	0.7	1346	21	7
Belize	434	90	344	79	2	0.5	21	5	2
Costa Rica	2427	1902	525	22	8	0.3	35	1	9
Panama	6819	862	5957	87	32	0.5	1604	24	41
Norway	9863	4322	5541	56	28	0.3	285	3	245
Germany	91308	72252	19056	21	152	0.2	987	1	2561

and overlaps, as well as repeated near-identical places were additional findings.

The necessary granularity for the kind of data examined here of course depends on the application. It may range from 100 – 1000 meters or even up to a few km, in which case it would still fall into the offered accuracy. In those cases, the issue raised above about inverse geocoding or overlapping places may take precedence over positional accuracy.

Correction of the data is yet an open issue and remains as future work. Only some of the errors can be detected on the dataset itself, even less directly corrected. Therefore a solution might be to include and merge other sources and deal with varying granularity, coordinates, and naming conventions (e.g., [15] or [17]). Another option would be to use OpenStreetMap for conflation [3]. For example, [12] show that accuracy in OSM is related to population density and that the mean deviation error for road junctions in German cities is below 10m.

For some corrections, additional data sources would be needed. However, this information often has a similar quantity and quality as the actual data or is not freely available. In short, if the data situation is already difficult, there is less availability of additional data to improve it. Therefore, even an improved measurement and identification of less reliable data would be helpful. We envision additional work on analysis and improved correction strategies. We hope the data process of GeoNames may become open in the future to facilitate further improvements.

We hope this work sheds some light on potential pitfalls and helps to better analyze or model uncertainty to support an informed use of the data. This paper should also be helpful in estimating the imprecisions for certain countries and adapting an error model accordingly. It might even stimulate improved analysis or motivate new correction strategies.

Acknowledgments

This paper was written during the tenure of an ERCIM "Alain Bensoussan" Fellowship Programme which has received funding from the European Union 7th Framework Programme (FP7/2007-2013), grant agreement n° 246016.

6. REFERENCES

[1] A. I. Abdelmoty and C. B. Jones. Towards maintaining consistency of spatial databases. In *CIKM'97*, 1997.

[2] D. Ahlers. Towards Geospatial Search for Honduras. In *LACNEM 2011*, 2011.

[3] D. Ahlers. Multi-source conflating index construction for local search in a low-coverage country. In *LA-WEB 2012*, 2012.

[4] D. Ahlers. Lo mejor de dos idiomas – Cross-lingual linkage of geotagged Wikipedia articles. In *ECIR2013*.

[5] D. Ahlers and S. Boll. On the Accuracy of Online Geocoders. In *Geoinformatik 2009*, 2009.

[6] R. Bakshi, C. A. Knoblock, and S. Thakkar. Exploiting Online Sources to Accurately Geocode Addresses. In *GIS '04*, 2004.

[7] P. V. Bolstad and J. L. Smith. Errors in GIS: Assessing Spatial Data Accuracy. *J. Forest.*, 90(11), 1992.

[8] T. J. Brunner and R. S. Purves. Spatial autocorrelation and toponym ambiguity. GIR '08, 2008.

[9] R. Devillers, A. Stein, Y. Bédard, N. Chrisman, P. Fisher, and W. Shi. Thirty years of research on spatial data quality: achievements, failures, and opportunities. *Transactions in GIS*, 14(4), 2010.

[10] D. W. Goldberg, J. P. Wilson, and C. A. Knoblock. From Text to Geographic Coordinates: The Current State of Geocoding. *URISA*, 19(1), 2007.

[11] C. Hauff. A Study on the Accuracy of Flickr's Geotag Data. SIGIR '13, 2013.

[12] M. Helbich, C. Amelunxen, P. Neis, and A. Zipf. Comparative Spatial Analysis of Positional Accuracy of OpenStreetMap and Proprietary Geodata. *Angewandte Geoinformatik*, 2012.

[13] L. L. Hill. Core Elements of Digital Gazetteers: Placenames, Categories, and Footprints. In *ECDL '00*, 2000.

[14] T. J. Holmes and S. Lee. Cities as six-by-six-mile squares: Zipf's law? *Agglomeration Economics*, 2010.

[15] H. Manguinhas, B. Martins, and J. L. Borbinha. A Geo-Temporal Web Gazetteer Integrating Data From Multiple Sources. In *ICDIM*, 2008.

[16] D. Potere. Horizontal Positional Accuracy of Google Earth's High-Resolution Imagery Archive. *Sensors*, 8(12), 2008.

[17] L. A. Souza, C. A. Davis, Jr., K. A. V. Borges, T. M. Delboni, and A. H. F. Laender. The Role of Gazetteers in Geographic Knowledge Discovery on the Web. In *LA-WEB '05*, 2005.

[18] J. Wieczorek, Q. Guo, and R. Hijmans. The point-radius method for georeferencing locality descriptions and calculating associated uncertainty. *Int. J. Geogr. Inf. Sci.*, 18(8), 2004.

[19] J. Zhang and M. Goodchild. *Uncertainty in Geographical Information*. Taylor and Francis, 2002.